Choosing the tree which actually best explains the data: another look at the bootstrap in phylogenetic reconstruction

V. Berry & O. Gascuel & G. Caraux
Département d'Informatique Fondamentale
LIRMM
161 rue Ada, 34392 Montpellier, France

1 Introduction

Phylogenetic Reconstruction aims at estimating the differentiation process resulting in a set S of n species (Swofford and Olsen, 1990). This process is usually represented by a valued tree t, called a *phylogeny* or *evolutionary tree*. This tree is usually considered as fully resolved, *i.e.*, as ternary. Let d be the tree distance induced on S by the path-lengths in t and let D denote the $n \times n$ distance matrix associated with d.

Here we suppose that t, and thus D, is unknown. Our aim is to retrieve the structure of the differentiation process, *i.e.*, the topology of t, on the basis of molecular data. For a fixed part of the genome we know the nucleotide subsequences associated with each species of S. We thus have a data matrix X consisting of n sequences of size p. Columns of X can be considered as a sample of homologous sites. If we knew the entire genome of the species S, then we could compute D from X, assuming that the evolution conforms to a known mathematical model, such as those of Jukes and Cantor (1969) and Kimura (1980). From the additive matrix D we could then retrieve the unique tree t with which it is associated (Sattah and Tversky, 1977; Barthélemy and Guénoche, 1991) and our goal would then be achieved. Knowing only a segment of the species genome, we can only compute an approximation of the unknown matrix D. This approximation $\hat{D}$ is used to estimate t by minimization of a certain goodness of fit criterion δ. This function measures the accuracy of $\hat{t}$ to represent D. Let τ denotes any valued tree whose leaves are associated with the species of S. $\delta(D, \tau)$ is minimum when $\tau = t$ and $\hat{t}$ is, therefore, defined as

$$\hat{t} = \underset{\mathcal{T}}{\operatorname{argmin}}\ \delta(\hat{D}, \tau). \qquad (1)$$

The tree $\hat{t}$ will generally be fully resolved. Indeed, adding any new ramification in the tree may only improve the value of the criterion. However, is the complete resolution of $\hat{t}$ relevant? Would the multiplicity of edges not lead to an overfitting effect? By seeking to adjust $\hat{D}$ as closely as possible do we not wander from D at some given step, inferring edges that lead to minimizing $\delta(\hat{D}, \hat{t})$ but to increasing $\delta(D, \hat{t})$?

If such were the case, a new sample of sites would very probably put at least a part of $\hat{t}$ topology into question. This potential variability raises doubts about the reliability of inferring fully resolved trees. For our purpose it is more interesting to search for less resolved, and thus more homogenous, topologies. In return, we have to accept a structural bias in terms of *complexity*, *i.e.*, number of edges. We then encounter a well-known problem in statistical estimation, that is, finding the best bias/variance compromise.

Searching for a less complex tree than $\hat{t}$ is connected with the problem of estimating the reliability of tree edges. Among other methods (see (Li and Gouy, 1990) for a survey), resampling techniques such as the bootstrap method (Efron, 1979; Efron and Tibshirani, 1993) are most commonly used.

Two strategies can be designed to find a tree of more appropriate complexity than $\hat{t}$. The first, qualified as ascending, consists in successively resolving a tree of null complexity (a star) until no new resolution is significant. On the other hand, one may adopt a descending strategy. A tree of maximal complexity (fully resolved) is first inferred and then submitted to an elimination process successively removing its edges (Fig. 1 shows how edges are removed).

The first strategy is computationally very expensive. It requires the examination of an exponential number of alternatives (bipartitions of species) at each step. The second strategy can be implemented more easily. Many algorithms are available for constructing fully resolved trees from a distance matrix (Sattah and Tversky, 1977; Saitou and Nei, 1987; Gascuel and Levy, 1996). We are interested in this particular strategy and investigate two approaches from this standpoint.

The first approach consists in removing from $\hat{t}$ the edges whose reliability is lower than some *a priori* fixed threshold. The number of removed edges thus depends on the chosen reliability threshold. Among other methods, Felsenstein (1985) introduced a now widespread bootstrap procedure. We showed recently (Berry and Gascuel, 1996) how the best reliability threshold can be chosen for this approach, according to some classical error measure.

The second approach consists in first determining the most appropriate complexity for the inferred tree, *i.e.*, the one leading to the best bias/variance

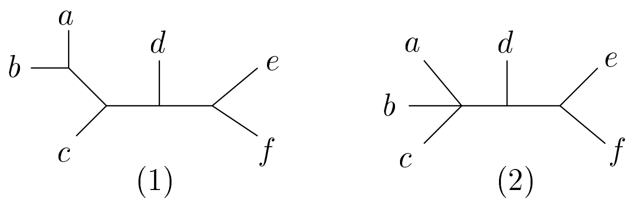

Figure 1: Removing the bipartition $\{a,b\}/\{c,d,e,f\}$

compromise. The edges of the tree $\hat{t}$ that is initially obtained are then sorted according to their importance in the minimization of $\delta(\hat{D},\hat{t})$. The less effective edges in this sense are then removed so that the tree reaches the determined complexity.

In the following we first define the distance criterion δ on the base of the *quartet* notion. Then (section 3) we describe a specific heuristic method to infer trees from distance matrices. Next (section 4,5) we detail the two descending approaches mentioned above and the resampling techniques on which they both rely. Finally (section 7) we report some simulation results comparing both approaches.

2 A topological goodness of fit criterion

In this study our only aim was to estimate the topology of a tree. Therefore, we did not consider the distance criteria usually employed in the phylogenetic reconstruction context (least squares, minimum evolution principle) since they firstly rely on edge lengths. Instead we chose a criterion which is a direct function of the tree structure. More precisely, it is based on quartets of species, *i.e.*, sets of four species.

Any dissimilarity D is a tree distance if and only if it satisfies the *four-point condition* (Buneman, 1974), also called the additivity condition. This condition requires that for any quartet of species $\{i,j,k,l\}$ the larger two of the three sums $D_{ij}+D_{kl}$, $D_{ik}+D_{jl}$, $D_{il}+D_{jk}$ be equal. If $D_{ij}+D_{kl}$ is the smallest sum, then there must be at least one edge separating i,j from k,l in the tree t representing D. This topological constraint is denoted $ij||kl$ and can be represented by a *4-tree*, as in Fig. 2(1). As can be seen from Fig. 2(1-3), four species can be separated into two pairs in three different ways. We denote these three situations respectively, $ij||kl$, $ik||jl$ and $il||jk$. When the three sums are equal, the quartet is unresolved. The latter situation is represented by a "star" on the four species (Fig. 2(4)).

In order to represent any distance matrix $\hat{D}$ in terms of 4-trees, Sattah and Tversky (1977) and Fitch (1981) suggest using the following relaxation

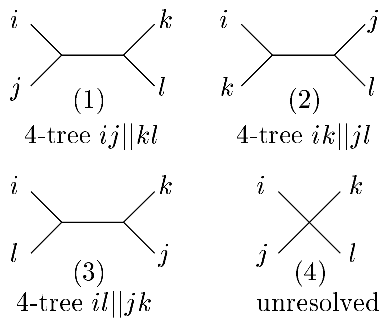

Figure 2: possible 4-trees and irresolution for the quartet $\{a, b, c, d\}$

to the 4-point condition:

$$\hat{D}_{ij} + \hat{D}_{kl} < \left\{ \begin{array}{l} \hat{D}_{ik} + \hat{D}_{jl} \\ \hat{D}_{il} + \hat{D}_{jk} \end{array} \right. \Leftrightarrow ij||kl. \tag{1}$$

This results from the fact that such a distance pattern is best approximated by the 4-tree of Fig. 2(1), *i.e.*, when i, j are separated from k, l. This was demonstrated for several quantitative goodness of fit criteria (Saitou and Nei, 1987; Gascuel and Levy, 1996). If none of the three sums in (1) is smaller than the others, no 4-tree is associated with the quartet $\{i, j, k, l\}$ (Fig. 2(4)).

Any tree topology τ on n species can also be defined by a 4-tree set (Bandelt and Dress, 1986; Steel, 1992). These 4-trees are the ones supported by its edges. More precisely, any edge e of τ separates S into two distinct parts, σ and σ', according to the two subtrees it defines in τ. If $\sigma|_e\sigma'$ denotes the bipartition induced on S by an edge e, then

$$\exists \sigma|_e\sigma' \in \tau \text{ s.t. } \{i, j\} \subset \sigma \text{ and } \{k, l\} \subset \sigma' \Leftrightarrow ij||kl \tag{2}$$

enables to associate a 4-tree set to τ. If the paths (ij) and (kl) in τ intersect in only one node, no 4-tree is associated with the quartet $\{i, j, k, l\}$. The tree τ can be considered as topologically equivalent to the 4-tree set obtained from (2).

Since any distance matrix and any tree can be associated with a 4-tree set, we will use this common concept to define the goodness of fit criterion δ. Robinson and Foulds (1981) defined the topological distance between two trees as the cardinal of the symmetrical difference between the sets of bipartitions induced respectively by their edges. Similarly, we define $\delta(\hat{D}, \hat{t})$ as the cardinal of the symmetrical difference between $\hat{D}$ and $\hat{t}$, expressed in terms of their associate 4-tree sets, $Q_{\hat{D}}$ and $Q_{\hat{t}}$, obtained from (1) and (2) respectively:

$$\delta_1(\hat{D}, \hat{t}) = |\{q \in Q_{\hat{t}} / q \notin Q_{\hat{D}}\}|,$$

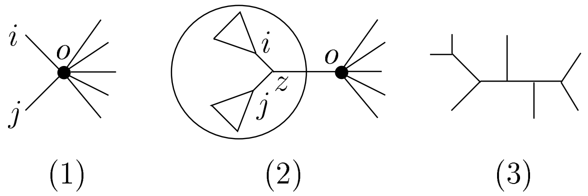

Figure 3: Obtaining $\hat{t}$

$$\delta_2(\hat{D},\hat{t}) = |\{q \in Q_{\hat{D}} / q \notin Q_{\hat{t}}\}|,$$
$$\delta(\hat{D},\hat{t}) = \delta_1(\hat{D},\hat{t}) + \delta_2(\hat{D},\hat{t}). \quad (3)$$

Thus δ measures the topological agreement of $\hat{t}$ and $\hat{D}$. δ can be applied to any distance matrix and therefore possibly to the true distance matrix D. In this case, if $\delta(D,\hat{t}) = 0$ then $\hat{t} = t$.

3 Constructing trees from quartets

Optimizing (3) is an NP-Hard problem. As usual in Phylogenetic Reconstruction, we rely on a heuristic method to infer the tree $\hat{t}$. The method is based on an agglomerative process (Fig. 3) as often in this context (Sattah and Tversky, 1977; Saitou and Nei, 1987; Gascuel and Levy, 1996). Starting off with a trivial tree structure, we iteratively cluster subtrees, *i.e.*, sets of species. More precisely, the star topology (Fig. 3(1)), having only one internal node O, is initially considered. At each step two neighbor nodes i and j of O are clustered (Fig. 3(2)). That is, the species included in the subtrees of respective root i and j, are clustered as opposed to the other species of S[1]. This clustering results in a new node z connected to O, i and j, and in the removal of edges (Oi) and (Oj). The edge (Oz) represents the new ramification induced in the tree. The construction process is halted when O has only three neighbors left, *i.e.*, when no further ramification can be introduced between the species of S (Fig. 3(3)).

At each step the cluster which implies the greatest improvement to the adjustment of $\hat{D}$ is chosen. However, for this purpose we use a slightly different criterion than the one defined before. Indeed, due to the heuristic agglomerative approach we chose, minimizing (3) at each step does not guarantee the recovery of the correct tree when processing additive data (examples are available on request). Therefore, to choose among the several possible

[1]initially, species are directly connected to O, the corresponding nodes being considered as trivial subtrees

clusters we use the following heuristic criterion (to be maximized):

$$\delta'(\hat{D}, \hat{t}) = \frac{1 + |Q_{\hat{D}}| - \delta_2(\hat{D}, \hat{t})}{\delta_1(\hat{D}, \hat{t})}. \tag{4}$$

The numerator of this expression is always positive. Whenever $\hat{D}$ is additive and t fully resolved, at each step there is at least one cluster that does not contradict any 4-tree of $Q_{\hat{D}}$. Thus $\delta_1 = 0$, implying $\delta' = \infty$. Maximizing (4) implies choosing a cluster at each step which satisfies this property. Finally, we obtain a fully resolved tree which satisfies $\delta_1 = 0$, since no 4-tree from $Q_{\hat{D}}$ has been contradicted. That is, the value of the complete tree is infinite for δ'. Since this tree is fully resolved, we also have $\delta_2 = 0$, and, therefore, $\delta = 0$. In other words, the algorithm always retrieves the tree which represents an additive data matrix[2].

When $\hat{D}$ is not additive, maximizing (4) leads to minimizing δ_1 and δ_2 and, therefore, to minimizing (3).

3.1 Obtaining a sequence of trees

We will see later that one of the descending approaches requires the inference of a sequence of decreasingly complex trees $[t_{max}, \ldots, t_0]$. In this expression t_{max} is the tree $\hat{t}$ inferred by the tree-building method described above and t_0 denotes the star topology. The other trees of the sequence are obtained by successively removing the branches of t_{max}. Moreover, we aim at producing a tree sequence such that each tree optimizes the goodness of fit criterion (3), but still respects its attributed complexity. This implies that the edges of t_{max} have to be removed in a specific order, not necessarily contrary to the order in which they were obtained since they were heuristically inferred. We rely on a step by step elimination process to produce the other trees of the sequence from t_{max}: its edges are reevaluated by the criterion (3). The edge whose elimination least decreases this criterion is then selected and removed. The next tree of the sequence is thus obtained. The elimination process is repeated until all edges of t_{max} have been removed, *i.e.*, t_0 has been obtained.

4 Fixed reliability threshold approach

As underlined before, there exist many methods to test the reliability of tree edges (Li and Gouy, 1990; Li and Zharkikh, 1995; Rzhetsky et al.,

[2]variants can be designed for dealing with additive matrices associated with non-fully resolved trees.

1995). The most widespread methods rely on resampling techniques, and more precisely on the bootstrap process (Efron, 1979; Efron and Tibshirani, 1993). Felsenstein (1985) proposed to resample the columns of the data matrix X with replacement to produce a new $n \times p$ data matrix X^* with the same number of sites. This resampling process aims at simulating the sampling variance with which the observed sites are obtained. These sites represent only a small part of the species sequence and are assumed to be independently obtained.

The same method as for X is used to infer a tree from X^*. We note $\hat{D}^*$ the distance matrix between species of S computed from X^*, and $\hat{t}^*$ the tree obtained from $\hat{D}^*$. According to (Efron et al., 1995) we denote this process

$$X^* \longrightarrow \hat{D}^* \longrightarrow \hat{t}^*.$$

This whole process is independently repeated a great number of times r. Then for each edge of $\hat{t}$ we compute the frequency (sometimes called the *bootstrap proportion*) with which it is inferred in the various trees $\hat{t}^*$ Each frequency f can be considered as a reliability measure for the corresponding edge. Edges inferred with low frequency will very probably be highly dependent on X and are doubtful to represent some topological aspect of the true tree t. These edges are very likely to result from the overfitting effect previously mentioned.

In introducing this use of the bootstrap technique in the phylogenetic reconstruction field, Felsenstein only aimed at estimating the reliability of a particular prespecified edge of $\hat{t}$. Here we use this procedure to estimate the reliability of all edges of $\hat{t}$. Edges are susceptible of being removed from $\hat{t}$ at the lower reliability levels. This implies statistical order problems to determine the exact validation threshold. Moreover the estimated reliability levels of the different edges are not independent. The test procedures based on the bootstrap technique appear to be reliable but of low power, because of this difficulties. However, as this area is still being thoroughly investigated (Zharkikh and Li, 1995; Efron et al., 1995), more effective procedures may be expected in the near future.

Using simulations (Berry and Gascuel, 1996) we have shown that a 95% threshold gives estimations of t that are too conservative. We have shown that the majority-rule tree (obtained from $\hat{t}$ with a threshold $f > 50\%$) is theoretically expected to be the best estimate of t according to the Robinson and Foulds (1981) distance measure. This is only so when assuming that f represents the probability that the corresponding edge be correct. As a matter of fact, f was shown by simulations (Berry and Gascuel, 1996) to be close to this probability, especially when $f \approx 50\%$.

Thus, reducing the overfitting effect by the above approach enables us to improve the quality of the estimation of t. However, a major defect of this approach is that the reliability threshold at which edges of $\hat{t}$ are removed is as yet chosen only empirically. Moreover, the best value for this threshold depends on many factors, such as the tree-building method and evolutionary conditions. It would appear that designing a theoretical expression to fix this threshold correctly will be problematic.

5 Variable complexity level approach

The second approach we present focuses on the most appropriate complexity level α for the tree $\hat{t}$. To avoid the previous drawback of arbitrary choice of a fixed threshold, we estimate α according to the reconstruction context, *i.e.*, the particular available data X and the tree-building algorithm.

For this purpose we rely on resampling techniques such as bootstrap and the cross-validation to produce "independent" data pseudosamples. An estimation $\hat{\alpha}$ of the ideal complexity is obtained by evaluating the bootstrap trees $\hat{t}^*$ with $\hat{D}$, inferred from the pseudosamples. In other words, we draw conclusions for $t - \hat{t}$ on the basis of $\hat{t} - \hat{t}^*$, which is the basic purpose of the bootstrap.

First a sequence $\hat{s}$ of decreasingly complex trees is inferred from $\hat{D}$:

$$X \longrightarrow \hat{D} \longrightarrow \hat{s} = [\hat{t}_{max}, \ldots, \hat{t}_0],$$

where $\hat{t}_{max} = \hat{t}$ and

$$\hat{t}_{max} \succ \ldots \succ \hat{t}_0,$$

The sequence $\hat{s}$ is obtained as described in section 3.1, *i.e.*, such that its trees are of decreasing efficiency in adjusting $\hat{D}$ in the sense of (3). This sequence of trees may be compared, to some extent, with the sequence of decision trees inferred in CART (Breiman et al., 1984), even if the purposes are different.

The original data matrix X is resampled, in the same way as previously described, to produce data pseudo-samples X^*. The corresponding distance matrix, $\hat{D}^*$, is computed each time and serves as a basis for obtaining a sequence $\hat{s}^*$ of trees of decreasing complexity:

$$X^* \longrightarrow \hat{D}^* \longrightarrow \hat{s}^* = [\hat{t}^*_{max}, \ldots, \hat{t}^*_0],$$

$$\hat{t}^*_{max} \succ \ldots \succ \hat{t}^*_0.$$

The trees of $\hat{s}^*$ are inferred by the same tree-building method as the one used to infer the trees of $\hat{s}$. Once $\hat{s}^*$ has been obtained, the accuracy of each of its trees to adjust the inital distance matrix $\hat{D}$ is measured, using (3). The best complexity $\hat{\alpha}^*$ for inferring a tree from X^* is deduced as being the one associated with the tree of $\hat{s}^*$ which best adjusts the topology of $\hat{D}$, *i.e.*, the tree $\hat{t}^*_{\alpha_i}$ such that $\hat{\alpha}^* = argmin_{\alpha_i}(\delta(\hat{D}, \hat{t}^*_{\alpha_i}))$.

The whole process leading from X^* to $\hat{\alpha}^*$ is repeated r' times. Remark that the same number of replications as for the fixed reliability threshold approach is not necessary. $r' = 10$ enabled to obtain sufficiently precise estimations, while $r = 1000$ is usually chosen for the other approach. An explanation is that here we aim at estimating only one value (the complexity α), whereas the previous approach aims at estimating as many values as the number of edges of $\hat{t}$.

The complexity $\hat{\alpha}$, used for the tree inferred from $\hat{D}$, is estimated as the mean of the $\hat{\alpha}^*$ values. Thus, the tree finally proposed to estimate t is the tree $\hat{t}_{\hat{\alpha}}$, *i.e.*, the tree of the sequence $\hat{s}$ which has $\hat{\alpha}$ as complexity.

Other resampling techniques can be used with the approach described above. For example, we also investigated the cross-validation technique. Its use necessitates some adjustments to the previous scheme. Each resampling consists in splitting the original data matrix X into two half-sized matrices X_1 and X_2. Columns of X are randomly attributed either to X_1 or to X_2. Let $\hat{D}_1$ and $\hat{D}_2$ be the two distance matrices computed from X_1 and X_2 respectively. $\hat{D}_1$ is used to produce the sequence $\hat{s}^*$ and $\hat{D}_2$ is used to evaluate its trees with the δ criterion (3). The complexity $\hat{\alpha}^*$ corresponds, as before, to the tree of $\hat{s}^*$ showing the best value in this sense. The process leading from X to $\hat{\alpha}^*$ is repeated r' times and $\hat{\alpha}$ is deduced as described before. $\hat{s}$ is obtained on the basis of all X and $\hat{D}$ matrices. The tree of $\hat{s}$ having $\hat{\alpha}$ as complexity level is chosen as the final estimation of t.

6 Simulations

We performed computer simulations to compare the relative efficiencies of the fixed threshold and the variable complexity approaches. For this purpose we simulated the evolution of molecular sequences along the edges of a model tree, as described in (Kuhner and Felsenstein, 1994). The same scheme of simulation as in (Berry and Gascuel, 1996) was used to produce tree estimates.

6.1 Generation of molecular data

First a rooted phylogeny t (topology and branch lengths) was randomly generated according to a poissonian process. Next a molecular sequence was randomly generated for the root (each of the four state $\{A, C, G, T\}$ having an equiprobability). The evolution of this sequence was simulated along the branches of t according to a Kimura 2-parameter model (Kimura, 1980) and to a preselected evolutionary condition. This condition specified whether the substitution rates were uniform among the various branches (or among characters). A data set X was then composed by the sequences obtained for the leaves of the tree.

We refer the reader to (Kuhner and Felsenstein, 1994) for more details on the simulation process. Programs used in these simulations are available on request.

6.2 Scheme of simulations

Starting off any data set X, three approaches are run in parallel:

- $\hat{D}$ is computed and given to the heuristic tree-building method of section 3, which results in the inference of a tree $\hat{t}$. This tree is a first estimate of the true tree t. We will refer to this approach as the *direct* one, noted (d), as opposed to the approaches based on resampling techniques.

- Resampling X provides data pseudosamples which are given as input to the fix threshold approach, (ft), described in section 4. We generated in this case $r = 100$ pseudosamples which is considered as a minimum for Felsenstein (1985) bootstrap method. We denote $\hat{t}_{ft}$ the estimate provided by this approach.

- The variable complexity approach (vc) also uses data pseudosamples. However in this case, $r' = 10$ replications were sufficient. Let $\hat{t}_{vc}$ denotes the estimate of t resulting from this approach, as described in section 5.

We compare average values for $\delta(D, \hat{t})$, $\delta(D, \hat{t}_{ft})$ and $\delta(D, \hat{t}_{vc})$, observed for various evolutionary conditions.

6.3 Results

Results showed that, in the context of the variable complexity approach, the cross validation technique is clearly less efficient than the bootstrap one. This may be due to the fact that it only uses half of the data set to construct the trees, and half of the data set to evaluate them. This leads to too conservative an approach. As a result, the cross validation estimated trees were always far less resolved than the other estimates of the true tree.

We report below results observed for $\hat{t}$, $\hat{t}_{ft}$ and $\hat{t}_{vc}$ (obtained with the bootstrap technique) on the basis of 20 trees $\times$ 20 data sets X. Each data set was obtained for 10 species using various conditions of evolution, such as constant substitution rates (c), unequal substitution rates between edges (u/b) and unequal substitution rates between sites (u/s). The conditions are presented in increasing order of difficulty. Displayed values represent error percentages computed in dividing $\delta(D, \hat{t})$, $\delta(D, \hat{t}_{ft})$ and $\delta(D, \hat{t}_{vc})$ by $|Q_{\hat{D}}|$, *i.e.*, by the number of data 4-trees. Table 1 and table 2 display average values observed for sequences of 100 sites and 300 sites respectively. With each mean value v is given in brackets its standard deviation, *i.e.*, $\sigma(v)/20$.

Table 1
Accuracy of the various approaches to estimate t (100 sites)

condition	(%) $\delta(D, \hat{t})$	(%) $\delta(D, \hat{t}_{ft})$	(%) $\delta(D, \hat{t}_{vc})$
(c)	36 (0.012)	8 (0.010)	14 (0.013)
(u/b)	36 (0.020)	22 (0.011)	25 (0.013)
(u/s)	34 (0.013)	30 (0.010)	32 (0.012)

Table 2
Accuracy of the various approaches to estimate t (300 sites)

condition	(%) $\delta(D, \hat{t})$	(%) $\delta(D, \hat{t}_{ft})$	(%) $\delta(D, \hat{t}_{vc})$
(c)	26 (0.015)	10 (0.010)	11 (0.013)
(u/b)	23 (0.010)	17 (0.008)	17 (0.007)
(u/s)	20 (0.007)	19 (0.010)	19 (0.005)

Both the (ft) and (vc) approaches appear well-founded since, in most cases, they lead to significant improvements in the results, with respect to the direct approach (up to 28% and 22% respectively, in terms of topological accuracy to the correct tree). The amount of the benefit greatly depends on the conditions of evolution. It reduces as the conditions grow more difficult. This simply originates from the fact that the (ft) and (vc) approaches rely on the same tree-building method that provides $\hat{t}$ and that this method encounters consistency problems for the most difficult conditions. More precisely, it does not always guarantee to find the correct tree even when having an infinite number of sites. As a consequence, all approaches provide in these cases less accurate estimates and the gap between $\hat{t}$ and $\hat{t}_{vc}$ or $\hat{t}_{ft}$ is reduced. Using sequences of higher size leads the bootstrap resampling to induce less variability between samples. Thus, the gap between the direct approach and the one relying on resampling is lowered for sequences of 300 sites.

The fixed threshold approach is often observed to be slightly more efficient than the variable complexity one, but the gap between them is usually small (at most of 6%, in terms of topological accuracy). On the other hand, the variable complexity approach is much faster to execute. Only 10 bootstrap replications are needed in order to have precise measures in this case. At least 100 replications are required for the fixed threshold approach.

The previously mentioned tendencies were also observed for various other number n of species.

7 Discussion

In phylogenetic reconstruction, usual tree-building method infer fully resolved trees which often contain some unreliable edges. Reducing the complexity (number of edges) of such trees was shown here to be an efficient strategy. More accurate estimates of the species history are obtained in removing certain branches of the fully resolved tree initially obtained.

We presented two different approaches relying on resampling techniques. The first approach fixes a reliability threshold and removes all tree edges showing a lower reliability level than this threshold. The other approach first estimates the most appropriate complexity and then removes the less reliable edges until the tree attain the desired complexity.

Fixing the threshold in the first approach can be delicate since only empirical methods exist for this purpose. Moreover, the optimal threshold depends on the conditions of evolution, which is not accounted for in the choice of the threshold. The second approach does not encounter this problem since the complexity is variable depending on the data set and on an underlying tree-building method. In addition, the variable complexity approach allows faster estimations than the first approach and provides almost as accurate estimates.

The variable complexity approach can also be applied to other criteria such as the least squares. Other parameters than the optimal complexity can be estimated by using the same approach. For example, we could aim at estimating the δ' threshold under which edges should be removed. When valued edges are of interest, the same approach could be used to determine the minimal length for edges to be kept in the tree... As can be seen, the approach presented in this paper can be used in many ways that deserve further investigations.

LITERATURE CITED

Bandelt, H.-J. and Dress, A. (1986). Reconstructing the shape of a tree from observed dissimilarity data. *Advances in applied mathematics*, 7:309–343.

Barthélemy, J. and Guénoche, A. (1991). *Trees and proximities representations.* Wiley.

Berry, V. and Gascuel, O. (1996). On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. *Mol. Biol. Evol.*, 13(7):999–1011.

Breiman, L., Friedman, J., Olshen, A., and C.J., Stone (1984). *Classification and regression trees.* Wadsworth.

Buneman, P. (1974). A note on metric properties of trees. *Journal of Combinatorial Theory*, 17(B):48–50.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statistics*, 7:1–26.

Efron, B., Halloran, E., and Holmes, S. (1995). Bootstrap confidence levels for phylogenetic trees. Technical Report Tech. Rep. 179, Standford University.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman et Hall.

Felsenstein85, J. (1985). Confidence Limites on Phylogenies : an approach using the bootstrap. *Evolution*, 39(4):783–791. (*Mesure de variabilite*).

Gascuel, O. and Levy, D. (1996). A reduction for approximating a (nonmetric) dissimilarity by a tree distance. In *Journal of Classification.*

Jukes, T. H. and Cantor, C. R. (1989). Evolution of protein molecules. In Munro, H. N., editor, *Mammalian Protein Metabolism*, volume III, chapter 24, pages 21–132. Academic Press, New York.

Kimura, M. (1980). A simple method for estimating evolutionary rates base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120.

Kuhner, M. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Comparaison de pleins de meths.

Li, W.-H. and Gouy, M. (1990). Statistical tests of molecular phylogenies. In Doolittle, R. F., editor, *Molecular evolution : computer analysis of protein and nucleic acid sequences*, volume 183 of *Methods In Enzymology*, chapter 40, pages 645–659. Academic Press.

Li, W.-H. and Zharkikh, A. (1995). Statistical tests of dna phylogenies. *Syst. Biol.*, 44:49–63.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.*, 53:131–147. Distance entre 2 topologies.

Rzhetsky, A., Kumar, S., and Nei, M. (1995). Four-cluster analysis: a simple method to test phylogenetic hypotheses. *Mol. Biol. Evol.*, 12(1):163–167.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstruction phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425.

Sattah, S. and Tversky, A. (1977). Additive similarity trees. *Psychometrika*.

Steel, M. (1992). The complexity of reconstructing trees from qualitative characters and subtrees. *J. of Classification*, 9:91–116.

Swofford, D. L. and Olsen, G. J. (1990). Phylogeny reconstruction. In Hillis, D. M. and Moritz, C., editors, *Molecular Systematics*, chapter 11, pages 411–501. Sinauer Associates, Inc., Sunderland, Massachusetts, USA.

Zharkikh, A. and Li, W.-H. (1995). Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. *Mol. Phylogenet. Evol.*, 4(1):44–63.